**Popis technického řešení datového uložiště PROJECT**

Technické řešení datového úložiště PROJECT obsahuje celkem tři identická souborová úložiště a další komponenty, které dohromady tvoří datové úložiště PROJECT. Souborová úložiště využívají interně technologii IBM Spectrum Scale / GPFS.

Souborové úložiště je vlastním NAS softwarově definovaným úložným řešením / produktem Dodavatele DATERA skládajícího se z komoditního serverového x86 hardware s operačním systémem Linux, blokových diskových polí škálovaných horizontálně pro zajištění požadovaného výkonu a kapacity, přístupové a LAN sítě. Softwarově jsou komponenty interně překryty pomocí IBM Spectrum Scale GPFS klastru a paralelním klastrovým souborovým systémem GPFS. NFS služby jsou poskytovány pomocí vysoce dostupného NFS sdílení, který ve vysoké dostupnosti exportuje data pomocí NFS protokolu.

1. **Komponenty Datového úložiště PROJECT**

Popis komponent a jejich konfigurací je uveden v následujících podkapitolách.

* 1. **Souborové úložiště**

Softwarově definované souborové úložiště jako ucelené NAS řešení poskytuje primárně pomocí NFS protokolu kapacitu požadovaným systémům zadavatele.

V podkapitolách této kapitoly jsou popsány parametry jednoho souborového úložiště.

* + 1. **GPFS Servery**

Každé souborové úložiště obsahuje redundantní GPFS servery, které tvoří základ klastru a do prostředí zákazníka jsou konfigurovány jako NFS servery poskytující vysoce dostupný NFS protokol. GPFS servery jsou realizovány pomocí x86 serverů, které jsou ke každému diskovému poli daného souborového úložiště připojeny napřímo pomocí redundantního 12Gbps SAS spojení, jsou redundantně zapojeny do vlastní LAN sítě za účelem správy a redundantně zapojeny do přístupové sítě za účelem zpřístupnění souborového úložiště.

Každé souborové úložiště obsahuje tři GPFS x86 servery, každý v konfiguraci:

* x 1x AMD EPYC2 Rome 7302P 3,0GHz 16-core CPU 128MB L3
* x 256GiB RAM 3200MHz DDR4 ECC
* x 2x 240GB SSD v RAID1 pro OS (hot-swap)
* x 16x 12Gbps mini-SAS HD portů
* x 1x Mellanox 100GbE dual-port QSFP28
* x Redundantní (hot-swap) zdroje včetně napájecích kabelů C13 - C14
* x Cable management rameno
* x RACK mount kit
* x IPMI 2.0 s KVM s vlastním 1GbE UTP LAN portem
* x Velikost 2U
* x Operační systém RedHat Enterprise Linux Server
  + 1. **Datové brány**

Každé souborové úložiště má dva x86 servery realizující datové brány, které mají redundantní vyhrazenou konektivitu do přístupové sítě. Tato konektivita je využita pro zpřístupnění souborového úložiště do vnitřní sítě a pro komunikaci s GPFS servery souborového úložiště. Jak pro připojení do sítě, tak pro připojení k serverům úložiště je poskytnuta agregovaná propustnost 100Gbps. Datové brány dále mají redundantní konektivitu do vlastní LAN sítě za účelem správy.

Parametry HW:

2x x86 server, každý o konfiguraci:

* 1x AMD EPYC Naples 7281 2,1GHz 16-core CPU 64MB L3
* 128GiB RAM
  + 8x 16GiB 2666MHz DDR4 ECC Registered 2R×8
  + AMD má 8 paměťových kanálů a moduly jsou tak rovnoměrně rozloženy
  + Teoretická propustnost:
    - Amd EPYC2 Rome (SPA3 LGA) 7302P - Per Socket Memory BW 204.8 GB/s
    - https://www.amd.com/en/products/cpu/amd-epyc-7302p
    - Amd EPYC Naples (SPA3 LGA) 7281 - Per Socket Memory BW 170.6 GB/s při použití DIMM 2666MHz, v kalkulaci 16GB 2666MHz DDR4 ECC Registered
    - https://www.amd.com/en/products/cpu/amd-epyc-7281
* 2x 240GB SSD v RAID1 pro OS (hot-swap)
* 1x Mellanox 100GbE dual-port QSFP28
* Redundantní (hot-swap) zdroje včetně napájecích kabelů C13 - C14
* Cable management rameno
* RACK mount kit
* IPMI 2.0 s KVM s vlastním 1GbE UTP LAN portem
* Velikost 1U
* Operační systém CentOS
  + 1. **Disková pole**

Souborové úložiště tvoří disková pole IBM FlashSystem 5030, která dohromady poskytují blokově kapacitu GPFS serverům skrze přímou SAS konektivitu. Pro správu jsou disková pole redundantně zapojena do vlastní LAN sítě souborového úložiště. Všechny GPFS servery daného souborového úložiště sdílí přístup ke všem diskům (LUNům) všech diskových polí daného souborového úložiště.

Parametry HW:

5x diskové pole IBM FlashSystem 5030, každé o těchto parametrech:

● 1x hlavní 2U řídící box:

○ Redundantní Active/Active hot-swap řadič, každý o těchto parametrech:

■ 1x 6-jádrový procesor

■ 32GB keš, zápisová keš chráněná baterií a zrcadlena mezi řadiči

■ 4x 12Gbps SAS porty pro připojení hostů

■ 2x 12Gbps SAS porty pro připojení diskových expanzí

○ Redundantní napájecí zdroj včetně kabelů C13/C14

○ 24x 2,5” pozicemi pro disky s dvěma redundantními SAS porty na disk

○ 19x 1,92TB 12Gbps SAS 2,5” Flash Drive

■ 1x Distribuovaný RAID6 8+2, rezervní hot-spare kapacita o kapacitě dvou disků

■ Užitečná kapacita 25,9TB / 23,6TiB

● 1x 5U HD expanze s 92x 3,5” pozicemi pro disky s dvěma redundantními SAS porty na disk

○ 68x 14TB 12Gbps SAS 3,5” 7,2k NL-SAS HDD

■ 2x Distribuovaný RAID6 8+2, každý o: 34x HDD a rezervní hotspare kapacitě o kapacitě dvou disků

■ Užitečná kapacita 707,4TB / 643,4TiB

○ Redundantní hot-swap napájecí zdroje včetně kabelů C19/C20

○ Redundantní SAS konektivita k řadičům

● IBM Spectrum Virtualize firmware

● mSAS HD kabely pro připojení hostů a expanse

2x diskové pole IBM FlashSystem 5030, každé o těchto parametrech:

● 1x hlavní 2U řídící box:

○ Redundantní Active/Active hot-swap řadič, každý o těchto parametrech:

■ 1x 6-jádrový procesor

■ 32GB keš, zápisová keš chráněná baterií a zrcadlena mezi řadiči

■ 4x 12Gbps SAS porty pro připojení hostů

■ 2x 12Gbps SAS porty pro připojení diskových expanzí

○ Redundantní napájecí zdroj včetně kabelů C13/C14

○ 24x 2,5” pozicemi pro disky s dvěma redundantními SAS porty na disk

● 1x 5U HD expanze s 92x 3,5” pozicemi pro disky s dvěma redundantními

SAS porty na disk

○ 68x 14TB 12Gbps SAS 3,5” 7,2k NL-SAS HDD

■ 2x Distribuovaný RAID6 8+2, každý o: 34x HDD a rezervní hotspare kapacitě o kapacitě dvou disků

■ Užitečná kapacita 707,4TB / 643,4TiB

○ Redundantní hot-swap napájecí zdroje včetně kabelů C19/C20

○ Redundantní SAS konektivita k řadičům

● IBM Spectrum Virtualize firmware

● mSAS HD kabely pro připojení hostů a expanze

Detailní informace o diskových polích lze získat zde:

https://www.ibm.com/products/storwize-v5000e

* + 1. **LAN síť**

Každé souborové úložiště má svou LAN síť skládající se ze dvou switchů Dlink DGS-1510-52X.

Funkčnost a konfigurace každého LAN switche:

* 48 x 10/100/1000Mbps
* 4x 10Gbit SFP+ uplink port
  + 1x 10Gbit uplink port k prvkům přístupové sítě
  + 2x 10Gbit - stack port
* Celková přepínací kapacita 176Gbps
* MAC Address Table - až 16384 per switch
* Port standardy / funkce - IEEE 802.3 10BASE-T Ethernet, IEEE 802.3u 100BASE-TX Fast Ethernet, IEEE 802.3ab 1000BASE-T Gigabit Ethernet, 802.3ae 10 GbE, IEEE 802.3x Flow Control for Full-Duplex Mode, Auto-negotiation
* Vzdálený management (https, ssh)
* 2x D-link DEM-CB100S SFP+ DAC kabel, 10Gbit, 1m pro propojení switche do stacku
* Velikost 1U
* RACK mount kit
* Napájecí a konzolový kabel
* Redundance napájení je realizována pomocí spravovatelného ATS switche.

Detailní informace o LAN switchích lze získat zde:

<https://eu.dlink.com/es/es/-/media/business_products/dgs/dgs-1510/datasheet/dgs_1510_series_datasheet_en_eu.pdf>

Každé souborové úložiště obsahuje jeden ATS switch, který napájí dva LAN switche. Je použit ATS switch s těmito vlastnostmi:

* Vstupní napájení: 2x C20
* Výstupní konektory napájení: 8x C13
* Vzdálená správa pomocí SSH

LAN sítě zajišťují komunikaci a správu zařízení uvnitř souborových úložišť, zejména:

* Komunikace mezi servery
* Management diskových polí
* Management serverů (IPMI)

Ve schématu níže je zachycena detailní topologie LAN sítě.

A diagram of a computer network

AI-generated content may be incorrect.

*Obrázek 1 - Schéma LAN sítě*

Servery mají redundantně zapojeny 1 GbE management interfacy do LAN switchů. Na úrovni OS Linux serverů jsou porty nastaveny do bondu pro dosažení redundance interfaců pro management LAN.

Disková pole mají redundantně zapojeny management interfacy do LAN switchů. Vysoká dostupnost managementu diskových polí je řešena nativně na úrovni diskového pole.

* + 1. **Technologické řešení souborového úložiště**

Klíčovým prvkem řešení souborových úložišť je technologie IBM Spectrum Scale / IBM GPFS.

Technologie IBM Spectrum Scale, která je v řešení využita, mimo jiné obsahuje:

* GPFS klastr
* Paralelní klastrový souborový systém GPFS
* ILM politiky pro řízení životního cyklu dat
* AFM pro asynchronní replikaci dat mezi GPFS klastry
* Vysoce dostupný NFS protokol
* CES klastr s NFS, SMB a objektovými protokoly
* GUI rozhraní pro správu včetně přehledu o kapacitách, výkonu a zátěže
* Performance monitoring
* Call Home funkcionalitu pro automatické hlášení problémů výrobci

Řešení má pro každé souborové úložiště s vlastním GPFS klastrem vytvořeným z třech GPFS serverů, které plní roli NSD, quorum, manager, GUI a NFS serverů. Klastr má redundancí N+1 na úrovni GPFS serverů.

V rámci tohoto klastru je vytvořen jeden souborový systém GPFS. V rámci tohoto systému souborů jsou nakonfigurovány dva GPFS pooly:

* System pool, ve kterém jsou na SSD discích uložena všechna metadata a pomocí ILM politik i malé soubory
* Datový pool, s daty větších souborů.

Díky jednomu souborovému systému GPFS vysdílenému skrze NFS protokol je zachován jednotný jmenný prostor celého souborového úložiště.

Detaily o IBM Spectrum Scale jsou k dispozici na těchto webových stránkách výrobce:

<https://www.ibm.com/support/knowledgecenter/en/STXKQY_5.0.3/ibmspectrumscale503_welcome.html>

<https://www.ibm.com/us-en/marketplace/scale-out-file-and-object-storage>

* + 1. **Přístupová síť**

Přístupová síť zajišťuje poskytování služeb datového úložiště systémům centra. Přístupová síť je tvořena dvěmi hraničními prvky CISCO Nexus 9336C-FX2.

Funkčnost a konfigurace každého hraničního prvku:

* 36 x 40/100-Gbps QSFP28 port
* Celková propustnost 7.2Tb/s
* Podpora Layer 3 (v4/v6) unicast a multicast routovacích protokolů včetně BGP, Open Shortest Path First (OSPF), Enhanced Interior Gateway Routing Protocol (EIGRP),
* Routing Information Protocol Version 2 (RIPv2), Protocol Independent Multicast Sparse Mode (PIMSM), Source-Specific Multicast (SSM) a Multicast Source Discovery Protocol (MSDP)
* Podpora RDMA over Converged Ethernet – RoCE
* Využitelný počet směrovacích záznamů směrovací tabulky hraničního prvku: 47100
* Využitelný počet oddělených routovacích instancí bez použití MPLS hraničního prvku: 1000
* Využitelný počet provozovaných VLAN hraničního prvku: 3967
* Využitelný počet provozovaných ACL záznamů hraničního prvku - vstupní pravidla: 5000
* Využitelný počet provozovaných ACL záznamů hraničního prvku - výstupní pravidla: 2000
* Počet volných portů 100Gbps Ethernet typu QSFP28 - hraniční prvek 1: 13
* Počet volných portů 100Gbps Ethernet typu QSFP28 - hraniční prvek 2: 13
* Podpora RADIUS , TACACS + a LDAP
* Redundantní napájecí zdroj
* RACK mount kit
* Velikost 1U

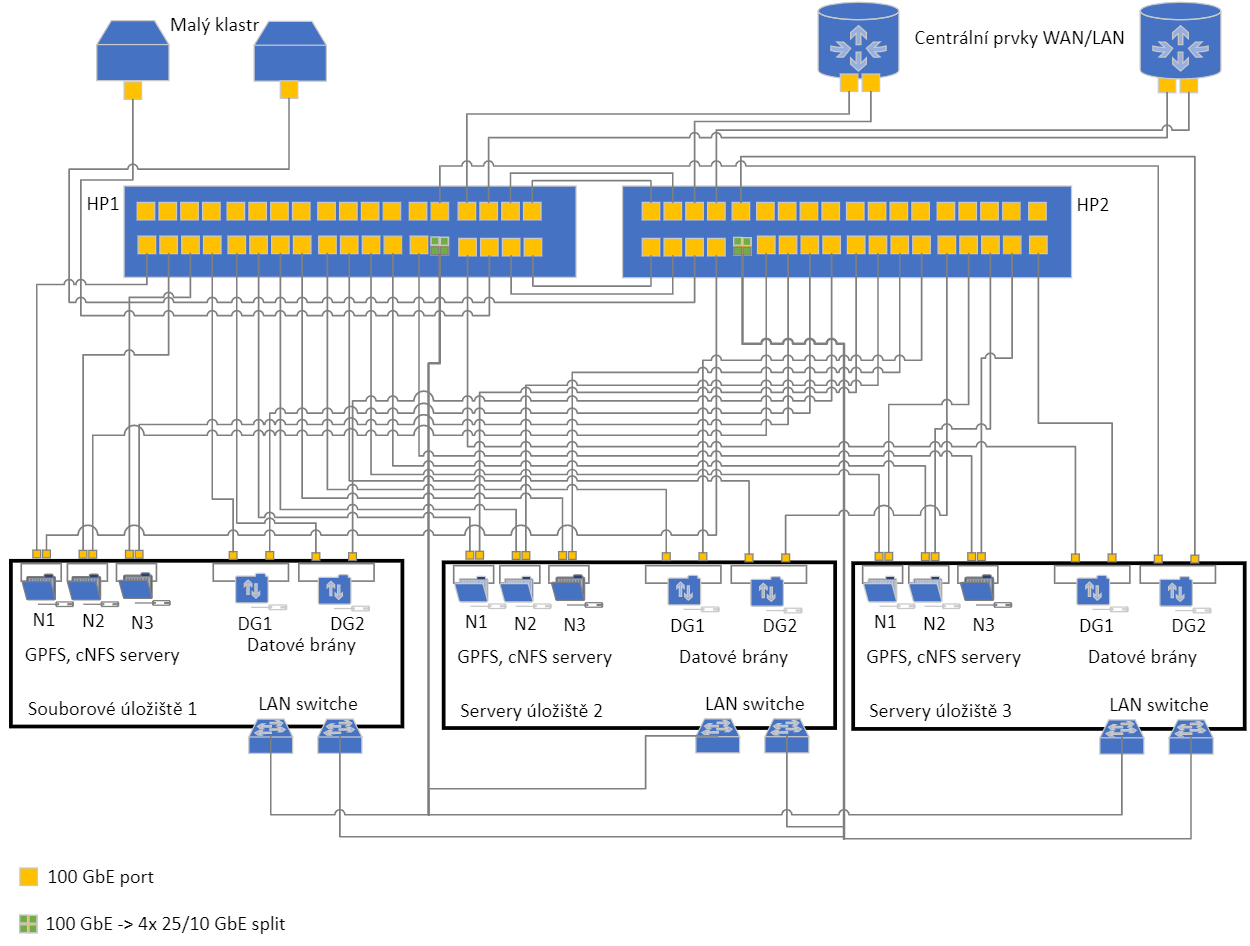
Switche CISCO Nexus 9336C-FX2 jsou včetně “LAN Enterprise License for Nexus 9300 Platform” licence.

Detailní informace lze získat zde:

<https://www.cisco.com/c/en/us/products/collateral/switches/nexus-9000-series-switches/datasheet-c78-742282.pdf>

Každý hraniční prvek je přes ethernet management rozhraní připojen do Out-of-band (OOB) sítě zadavatele.

Každý hraniční prvek poskytuje 13 volných 100 Gbps Ethernet portů typu QSFP28. Ve schématu níže je zachycena detailní topologie přístupové sítě.



*Obrázek 2 - Schéma přístupové sítě*

Pro přístup k primární NFS službě souborového úložiště je každé souborové úložiště do přístupové sítě připojeno pomocí šesti 100 Gbps Ethernet propojů. Vždy dva 100 Gbps Ethernet propoje z každého NFS serveru. Nad těmito propoji jsou vždy pro každý NFS server konfigurovány agregace.

Propoj mezi hraničními prvky je zajištěn pomocí čtyř pasivních kabelů QSFP-100G-CU1M “Cisco 100GBASE-CR4 QSFP Passive Copper Cable, 1-meter”.

Připojení do WAN/LAN sítě zadavatele je zajištěno moduly 4 x QSFP-100G-SR4-S=. V každém hraničním prvku 2x 100GBASE SR4 QSFP Transceiver, MPO, 100m over OM4 MMF, MPO-12 (12 fibers).

Downlink k LAN switchům je řešen pomocí dvou pasivních breakout kabelů QSFP-4SFP10G-CU5M ”Cisco 40GBASE-CR4 QSFP to 4 10GBASE-CU SFP+ direct-attach breakout cable, 5-meter, passive”. V každém hraničním prvku je jeden breakout kabel, který je využit pro připojení uplink portů pro 3x LAN switche.

Spojení s Malým klastrem je řešeno pomocí dvou aktivních optických kabelů QSFP-100G-AOC30M “Cisco 100GBase QSFP Active Optical Cable, 30-meter”. Vždy z každého hraničního prvku jeden propoj do serveru v Malém klastru.

Připojení serverů souborových úložišť (NFS i datových bran) je řešeno pomocí 100 Gbps metalických pasivních kabelů “QSFP-100G-CU5M”.

1. **Infrastruktura DC**
   1. **RACKy**

Celé řešení Souborových uložišť je umístěno do 5 RACKů.

5x RACK ATOS R@ck’n Roll 1200 42U RACK májí tyto parametry:

● Rozměry v mm: 2020v x 600š x 1400h

● Hmotnost RACKu s dveřmi 300Kg

● Přední perforované dveře

● Zadní aktivně vodou chlazené dveře

● 2x Plastová rozvodná skříň na střeše RACKu s přívodem EC60309 32A/5p

pro připojení jednotlivých jednofázových PDU k 3 fázové napájecí síti. PDU

se zásuvkami C19 a C13, polovina PDU připojena k jedné napájecí větvi (A),

druhá polovina ke druhé (B)

● Bočnice

● Ucpávky a kryty

Rozložení technologií do jednotlivých racků je zachyceno ve schématu níže.

A table of data with numbers

AI-generated content may be incorrect.

*Obrázek 3 - Rozložení technologií do racků*

A blueprint of a building

AI-generated content may be incorrect.

*Obrázek 4 - Schéma umístění RACKů v datovém centru*

* 1. **Chlazení**

Chlazení celého řešení je studenou vodou chladičem integrovaným v zadních dveřích každého RACKu.

* 1. **Napájení**

Každý z pěti RACKů je vybaven PDU umožňující redundantní připojení jednotlivých zařízení. Polovina redundantních zdrojů veškerých zařízení (s výjimkou LAN přepínačů) je zapojena do PDU připojených k jedné napájecí větvi (A), druhá polovina do PDU připojených ke druhé napájecí větvi (B). LAN přepínače, které mají jen jeden napájecí zdroj, jsou zapojeny do ATS přepínače, který je redundantně PDU připojených k různým napájecím větvím.

Každý RACK obsahuje:

● PDU napájená ze dvou napájecích větví A a B, každá větev:

○ 1x Přívod EC60309 3f 32A/5p

○ PDU se zásuvkami C19 a C13

1. **Schéma Datového úložiště**

Schématický nákres jednoho souborového úložiště je uveden na následujícím schématu.

Diagram of a diagram of a system

AI-generated content may be incorrect.

*Obrázek 5 - Schéma jednoho souborového úložiště*

Totožným způsobem jsou realizovány všechna souborová úložiště. Souborová úložiště jsou připojena do přístupové sítě.